

Deep Learning Report

Comparative Analysis of Vision Transformers and Convolutional Neural Networks for MRI Tumor Detection and Classification

Tumor-Classification-MA7CY070

Anonymous Student

May 5, 2025

Abstract and Introduction

Deep learning has revolutionized computer vision, with Convolutional Neural Networks (CNNs) being the cornerstone of many breakthroughs. Recently, Vision Transformers (ViT) have emerged as a promising alternative. This report explores the differences, advantages, and limitations of these two architectures in the field of medical imaging, with a focus on MRI tumor detection—a critical area in medical diagnostics. It provides a comparative analysis of their unique features and performance metrics.

1 Convolutional Neural Networks (CNN)

CNNs are a class of deep learning models designed to process structured grid data, such as images. They have been the standard for image classification tasks for years. Key components of CNNs include:

- **Convolutional Layers:** Apply filters to the input image (layered tensor) to detect patterns such as edges, textures, and shapes.
- **Pooling Layers:** Reduce the spatial dimensions of feature maps, improving computational efficiency and robustness to spatial variations.
- **Fully Connected Layers:** Combine extracted features to make predictions.

Depthwise Separable Convolution

Depthwise Separable Convolution is an efficient variant of the standard convolution operation. It splits the convolution into two steps:

- **Depthwise Convolution:** Applies a single filter to each input channel independently, extracting spatial features.
- **Pointwise Convolution:** Uses a 1×1 filter to combine the outputs of the depthwise step, reducing parameters and computations.

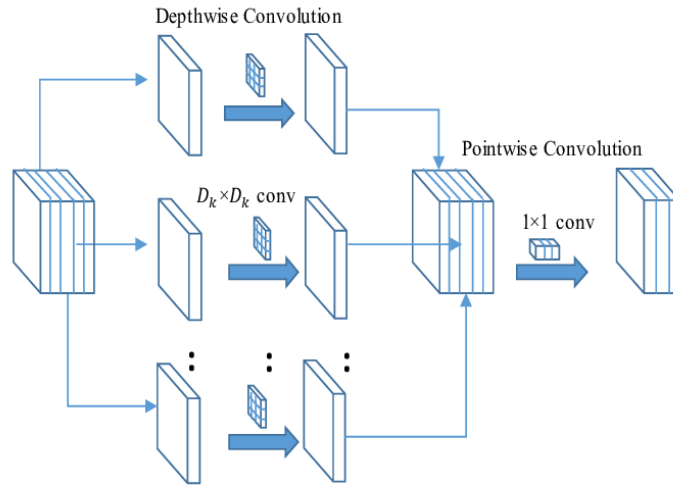


Figure 1: Illustration of Depthwise Separable Convolution Layer. Source.

EfficientNet and MBConv Blocks

EfficientNet is a family of convolutional neural networks designed for high accuracy *per* amount of computation. Its key ideas are:

- **Balanced Scaling:** Instead of increasing network depth, width, or input size in isolation, EfficientNet uses a single factor ϕ to scale all three together:

$$\text{depth} \propto \alpha^\phi, \quad \text{width} \propto \beta^\phi, \quad \text{resolution} \propto \gamma^\phi,$$

where α, β, γ are chosen so the model grows smoothly and efficiently.

- **MBConv Block:** The core building block, called Mobile Inverted Bottleneck Convolution, works in three steps:
 1. *Expand*: a 1×1 convolution increases channels from c to tc .
 2. *Depthwise*: a separate spatial filter per channel captures features with minimal cost.
 3. *Project*: another 1×1 convolution reduces back to c channels.

A skip (residual) connection adds the block’s input when shapes match, easing training.

- **Squeeze-and-Excitation (SE):** Optionally inside each MBConv, a tiny “attention” module:
 - Squeezes each channel to one value by global average pooling.
 - Learns per-channel weights via two small fully-connected layers.
 - Re-scales channels, letting the network focus on the most important features.

Putting it together, an EfficientNet model is just a stack of these MBConv blocks, scaled uniformly by ϕ . This yields networks (B0–B7) that achieve state-of-the-art accuracy with far fewer parameters and FLOPs than many traditional architectures.

We will not re-implement EfficientNet from scratch but use the provided EfficientNet from torchvision as a reference goal for the regular CNN and ViT models.

2 Vision Transformers (ViT)

Vision Transformers leverage the self-attention mechanism to process image data. Unlike CNNs, they do not rely on convolutional layers. The image is divided into patches, which are then treated as sequences of tokens with positional information, similar to words in natural language processing, as seen in the BERT (encoder-only transformer) architecture.

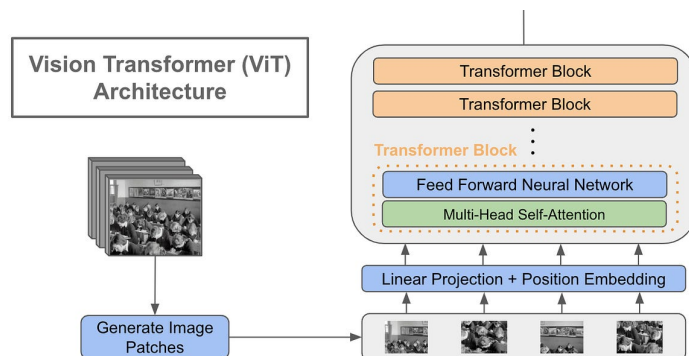


Figure 2: Illustration of Vision Attention Transformer. Source.

- **Patch Embedding:**

- The input image is divided into fixed-size patches (e.g., 16×16 pixels). They can be randomly sampled or uniformly spaced.
- Each patch is flattened into a vector and linearly projected into an embedding space.

- **Positional Encoding:**

- Since transformers do not inherently capture spatial information, positional encodings are added to the patch embeddings.
- These encodings help the model understand the relative positions of patches in the image.

- **Advantages:**

- Captures long-range dependencies and global context effectively.
- Scales well with large datasets and high computational resources.

- **Challenges:**

- Requires large datasets for training to achieve competitive performance.
- Computationally expensive compared to CNNs, especially for high-resolution images.

3 Expected Comparison

Below are key differences expected between the architectures based on the literature:

1. Inductive Bias

- **CNNs:** Strong spatial inductive bias (localized filters, hierarchical feature detection).
- **ViTs:** Weaker inductive bias, relying on self-attention to capture global dependencies.

2. Feature Detection

- **CNNs:** Detect local patterns (edges, textures) via convolutional filters.
- **ViTs:** Model long-range relationships through attention mechanisms.

3. Data Requirements

- **CNNs:** Perform well with limited data due to their hierarchical structure.
- **ViTs:** Require large datasets for pretraining to avoid overfitting.

4 MRI Brain Tumor Detection/Classification

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique widely used in medical diagnostics, particularly for brain imaging. Unlike CT scans, which use ionizing radiation, MRI employs strong magnetic fields and radio waves to produce detailed images of soft tissues. This makes MRI the preferred choice for detecting brain tumors, as it provides high-resolution images without exposing patients to harmful radiation.

Why MRI Tumor Detection?

Brain tumors are life-threatening conditions that require early and accurate diagnosis for effective treatment. MRI tumor detection aims to identify the presence of tumors and classify them into specific types. This classification is critical for determining the appropriate treatment plan, such as surgery, radiation therapy, or chemotherapy. Moreover, automated tumor detection systems can assist radiologists by reducing diagnostic time and improving accuracy.

MRI is particularly effective for capturing brain tissues because it detects the amount of hydrogen in different tissues, providing detailed contrast between soft tissues. In contrast, CT scans capture a "shadow" of density, which is less effective for distinguishing fine details in brain structures. However, MRI scans are longer and more challenging to perform, requiring patients to remain still for extended periods, which can be difficult in certain cases.

Dataset and Classification Problem

Usually, MRI images are 3D in DICOM/NifTI volume slices format, but for this study, we will use normalized 2D slices of the brain MRI images. The datasets consists of brain 2D MRI images categorized into four classes:

- **Glioma:** A type of tumor originating from glial cells in the brain and spinal cord.



Figure 3: Glioma Tumor

- **Meningioma:** Tumors arising from the meninges, the protective membranes surrounding the brain and spinal cord.

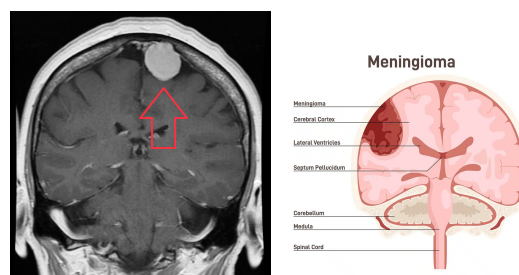


Figure 4: Meningioma Tumors

- **Pituitary:** Tumors developing in the pituitary gland, which regulates hormone production.

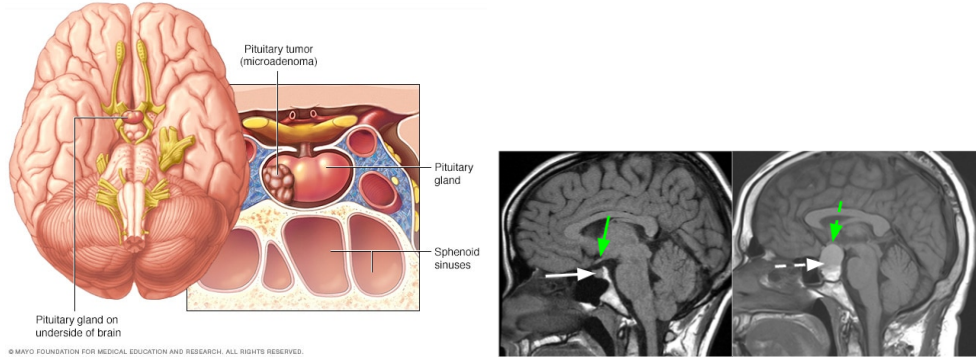


Figure 5: Pituitary Tumors

- **NoTumor:** Images representing healthy brain scans with no signs of tumor presence.

Datasets and Experimental Setup

For this study, we will use three publicly available datasets to train and validate our models. The datasets are as follows:

- **masoudnickparvar/brain-tumor-mri-dataset:**
 - Classes: Glioma, Meningioma, Pituitary, NoTumor.
 - This dataset will be used for training and primary validation.
- **sartajbhuvaji/brain-tumor-classification-mri:**
 - Classes: Glioma, Meningioma, Pituitary, NoTumor.
 - This dataset will be used for generalized validation over the classification task.
- **preetviradiya/brian-tumor-dataset:**
 - Classes: Brain Tumor, Healthy.
 - This dataset will be used for generalized validation over the tumor detection task (binary classification: tumor vs. healthy).

We will train a model to classify tumors using the first dataset and evaluate its generalization capacity on the other two datasets. This approach ensures that the model is not only effective in classification but also robust in detecting tumors across different datasets.

5 Results and Discussion

Benchmark Results

Validation Loss/Accuracy on the Original Dataset

The models were evaluated on the original dataset, and the results are as follows:

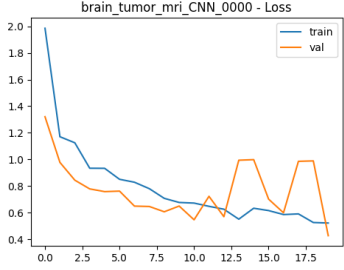
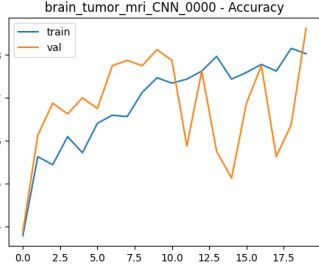
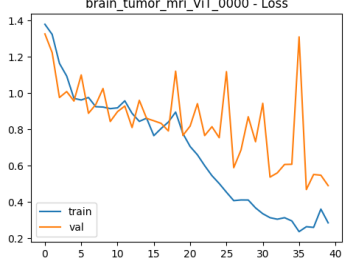
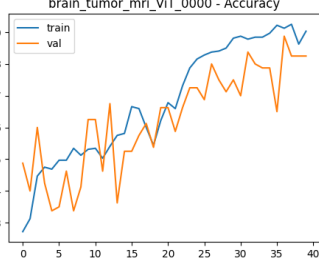
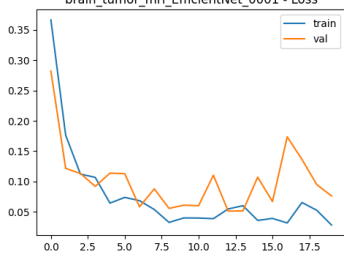
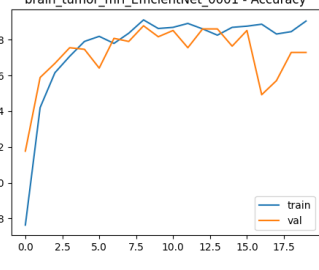
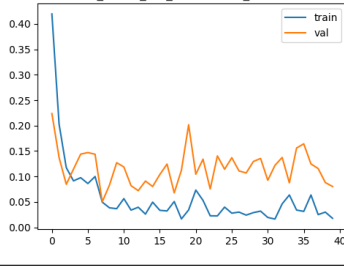
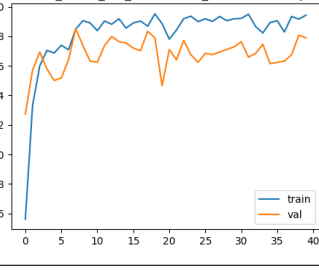
Model	Loss	Accuracy	Training Image	
CNN	2.40	39.05%		
ViT	2.92	55.00%		
EffNet-1	0.12	96.64%		
EffNet-0	0.10	97.33%		

Table 1: Comparison of Loss, Accuracy, and Training Images for CNN, EffNet-0 (low resolution), EffNet-1 (high resolution), and ViT.

Observation: EfficientNet (low and high resolution) models significantly outperform both the standard CNN and ViT in terms of accuracy and loss. However, ViT demonstrates better generalization compared to the standard CNN.

Validation on a Similar Dataset from a Different Source

The models were further validated on a similar dataset from a different source:

Model	Test Loss	Test Accuracy
Regular CNN	4.4039	0.2157
Vision Transformer	5.2470	0.3350
EfficientNet 1 (high resolution)	1.6221	0.7437
EfficientNet 0 (low resolution)	1.4960	0.7589

Table 2: Validation Results on a Similar Dataset from a Different Source

Observation: EfficientNet continues to outperform the other models, though its accuracy performance drops on this dataset (-21 percent). ViT shows better generalization than the standard CNN but still significantly lags behind EfficientNet.

Validation on a Third Dataset (Binary Classification)

For the third dataset, which involves binary classification (tumor vs. healthy), the results are as follows:

Model	Test Accuracy
Regular CNN	45.24%
Vision Transformer	50.09%
EfficientNet 1 (high resolution)	42.59%
EfficientNet 0 (low resolution)	54.54%

Table 3: Validation Results on the Third Dataset (Binary Classification: Tumor vs. Healthy)

Observation: On this binary classification task, Vision Transformers (ViTs) perform comparably to EfficientNet, showcasing their ability to generalize effectively (even with the low amount of data we have).

6 Conclusion

The benchmark results indicate that Vision Transformers (ViTs) perform and generalize better than standard CNNs. However, due to the limited amount of data available, their performance does not reach its full potential. Interestingly, EfficientNet, which represents an optimized choice of parameters for CNNs, outperforms ViTs on the current datasets and succeed to achieve similar result on way different data. This suggests that further investigation is needed to determine if the ViT architecture can be improved to achieve performance comparable to EfficientNet, even with the current data constraints.

While this study focuses on tumor classification, the ultimate goal is to extend these techniques to tumor segmentation. Accurate segmentation would enable precise localization of the tumor, aiding in surgical planning and treatment monitoring. The performance of the classifier will serve as a foundation for developing more advanced segmentation models.